ICC 2017
WASHINGTON
JULY 02-07, 2017

Research on a Method of Geographical Information Service Load Balancing

Li Heyuan[1] Li Yongxing[1] Xue Zhiyong[1] Feng Tao[1]

1. Xi'an Research Institute of Surveying and Mapping, Xi'an, Shaanxi, China; 305789861@qq.com; yongxingli2017@163.com; zhiyongxue2017@163.com; 1025019457@qq.com

Abstract：With the development of geographical information service technologies, how to achieve the intelligent scheduling and high concurrent access of geographical information service resources based on load balancing is a focal point of current study. This paper presents an algorithm of dynamic load balancing. In the algorithm, types of geographical information service are matched with the corresponding server group, then the RED algorithm is combined with the method of double threshold effectively to judge the load state of serve node, finally the service is scheduled based on weighted probabilistic in a certain period. At the last, an experiment system is built based on cluster server, which proves the effectiveness of the method presented in this paper.

Key words: geographical information service, load balancing, intelligent scheduling, concurrent access

1. Introduction

Geographical information is playing an increasingly important role on government management decision-making, national defense safety and people's living standards improve. Along with our country economy and the advancement of national defense informatization construction, agencies at all levels and the public's demand to authority, reliable geographic information service increases, need to implement comprehensive utilization and online services of multi-scale, multi-type of geographic information resources(Chen et al. 2009).The current network geographic information service system is facing the unpredictable challenge of concurrency growth, capacity limits and system response, and geographic information service scalability and availability is being more and more attention(Chen et al. 2013).Geographic information service dynamic load balancing is to solve how to make virtual multiple service node based on the Web server cluster into a logically unified "super" geographic information service center, implement the virtualization management and intelligent scheduling for each service node within the cluster resources, in order to support the service of dynamic binding, find and replace, thus improve the concurrent access ability of cluster system. Such as amazon services can be on-demand intelligent scheduling system capacity (such as servers, storage, and network bandwidth), can be flexible deployment a variety of services from provider resources, without the need for extra configuration for the uncertain demand in advance(Wang et al. 2010). Among existing network Map service systems, the Google Map has the intelligent scheduling ability of geographic information service resources with the support of dynamic load balance. In the field of other Web services, common load balance algorithm are rotation scheduling algorithm (RR) and local perception request distribution algorithm (LARD). The main problems of the RR algorithm is not considering the situation of load each service node, resulting in a decline in performance of the system. If a certain type of service request rate is high, the algorithm of LARD leads to a node of the utilization rate is very high, and the other nodes are idle for a long time, wasting server resources(Genova and Christensen 2000; Ren et al. 2010).

In this paper, we proposed a dynamic load balancing algorithm for multiple service node. According to service node in the cluster, using designed load balancing algorithm to share a large number of concurrent access to business

to multiple processing nodes respectively, reducing the time for a response, so as to achieve more service node's geographical information service virtualization and global load balancing.

2 General idea of the algorithm

The basic process of dynamic load balancing algorithm includes: receiving concurrent requests of users, load information of server collection feedback on the first task scheduler, determining the server load condition, selecting the best target server handling user requests to access concurrently. This algorithm adopts the centralized scheduling policy, the geographic information service node real-time monitoring their server load information, and feedback to the front-end server at a certain period, and then based on threshold method and RED algorithm for determining the node load state and obtain the node dynamic residual capacity. If the node is overload, there will be an effective warning. By classifying the geographic information service (conventional geographic information service types are divided into map browsing, query, analysis, data subscription and download, and metadata directory service, etc.), match a certain type of geographic information service to the specific application server group, on the basis of the residual capacity of the node dynamically determining the scheduling weights of multiple nodes. The node remaining capacity is proportional to the scheduling weights, and the node load current is inversely proportional to the scheduling weights, finally determine the optimum target server handling user requests to access concurrently. The application of model diagram is shown in Fig 1.

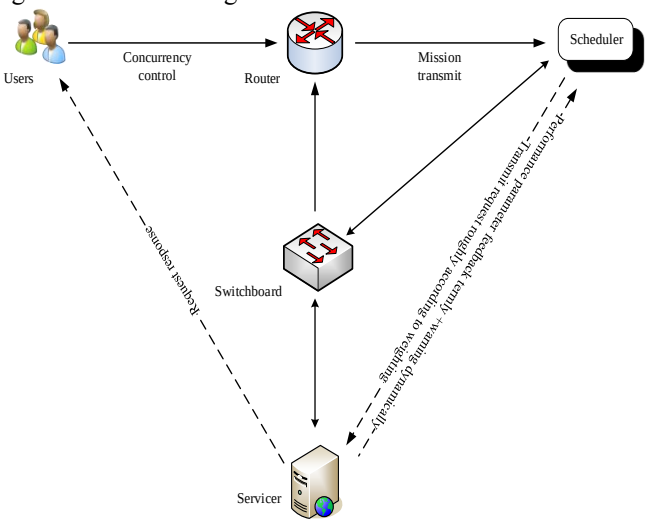

Fig. 1. Dynamic Load Balancing Algorithm Model of Geographic Information Service

When user concurrent access to a geographical information service, load balance algorithm comprehensively considers the load ability and the real-time service value, containing four server parameter, respectively is: the CPU usage, memory usage, network bandwidth utilization and disk I/O utilization. Service request via a router forwarding to the task scheduler, the scheduler collects load information at a certain cycle, and on the basis of the weighted value of load factor to determine the best target server. The target server will respond then service content is disseminated through switches and routers directly to the user.

In one cycle, a geographical information service for the user request is *S*, in the process load balance can be achieved by the following steps, as is shown in Fig 2.

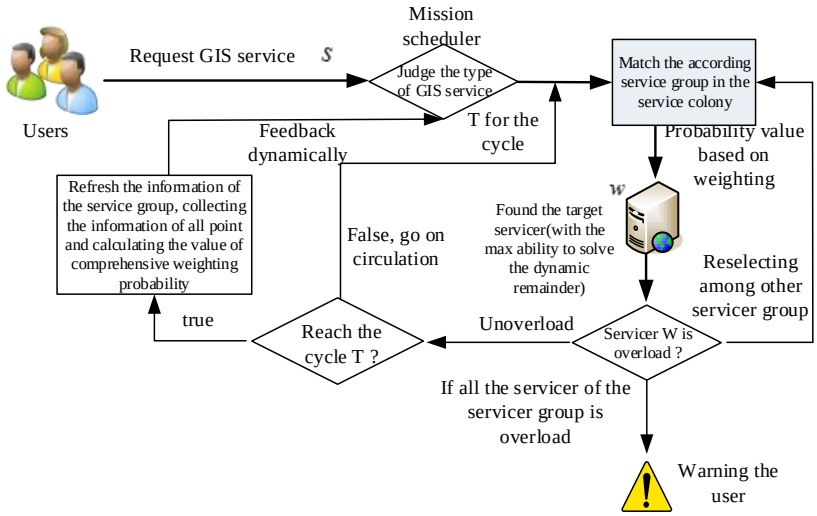

Fig. 2. Steps of Dynamic Load Balancing Algorithm

3 The specific design of the algorithm

3.1 Load information collection and real-time load calculation

Load information collection includes geographic information service type, the node work ability and the real-time load. The load information for the dynamic feedback cycle is T. This algorithm adopts the distributed load information collection, namely the cluster service node is monitoring dynamic real-time load information, and then dynamic feedback to the front-end periodic task scheduler.

$W=\{w_1,w_2,...,w_m\}$ s defined as the cluster server, $w_i(1\le i\le m)$ is defined as the number i server nodes in a cluster, m is defined as node number.

Server w_i current load $L_i=(L_{cpu},L_{I/O},L_{mem},L_{net})$, $L_{cpu},L_{I/O},L_{mem},L_{net}$ are defined respectively as the current CPU usage, disk I/O usage, memory usage and network bandwidth utilization rate of current service node.

Server w_i processing power $C_i=(C_{cpu},C_{I/O},C_{mem},C_{net})$, $C_{cpu},C_{I/O},C_{mem},C_{net}$ are defined respectively as service node rate of CPU, disk I/O speed, memory capacity and network throughput.

Geographic information service type collection $S=\{s_1,s_2,s_3,s_4\}$, $s_j(1\le j\le 4)$ is defined as the number j geographic information service type, among them, s_1,s_2,s_3,s_4 are defined respectively as map browsing, query analysis, data subscription and download, metadata and directory services.

When the same server support different types of geographic information services, it has the different effect on its load information [5].According to the differences in different geographic information service type occupies system resources, set a weight vector for each kind of geographic information service type：

$a_j=(a_j^{cpu},a_j^{I/O},a_j^{mem},a_j^{net})$, $|a_j|=1$, $j\in[0,3]$, a_j has the highest weight component，said that the geographic information service the more dependence on the corresponding resources.

Set server w_i provides services of type s_j, C_i as the node w_i processing power, L_i as the w_i load of utilization, C_m as a benchmark server processing ability, a_j as the corresponding weight vector to represent geographic information service s_j, the server w_i of the real-time load ratio value ω_i mbe defined as follows:

$$\omega_i = a_j^{cpu} \times \frac{C_i^{cpu} \times L_i^{cpu}}{c_m^{cpu}} + a_j^{I/O} \times \frac{C_i^{I/O} \times L_i^{I/O}}{c_m^{I/O}} + a_j^{mem} \times \frac{C_i^{mem} \times L_i^{mem}}{c_m^{mem}} + a_j^{net} \times \frac{C_i^{net} \times L_i^{net}}{c_m^{net}}$$

For introduce the benchmark server C_m, because in a heterogeneous environment, the working ability of different servers are need to be normalized. Through the formula 1, when ω_i is larger, that is to say the load of server is greater.

3.2 Determine the load condition based on the RED algorithm

The judgement of load state refers to the load value according to the service node, concludes that the node is idle or overload. This article uses the RED algorithm to determine the load of service node state, in order to avoid the happening of load dithering phenomenon (load jitter refers to the service node is marked as overload state, lead to other server nodes which will take charge for the additional load). RED (Random Early Detection) is put forward in packet-switched networks at first, to reduce congestion phenomenon in the greatest degree. The basic idea for determining the load status based on the RED algorithm is:

(1) If the load of service node value is lower than the threshold L_{low}, the node will be marked as overload state probability value of 0;

(2) If the load of service node values higher than the threshold L_{high}, the node will be marked as overload state probability value of 1;

(3) If the load of service node values are between L_{low} and L_{high}, the node will be marked as the state of the overload probability value of between 0 and 1, rather than absolutely marked as free or overload condition, the server at this time of real-time load is between L_{low} and L_{high}.

Set the server node w_i which is marked as overload has the state probability P, node load ratio value of real-time is ω_i, then the formula may be defined as follows:

$$P = \frac{k(\omega_i - L_{low})}{(L_{high} - L_{low})}$$

Among them, k is constant, L_{low} and L_{high} are the minimum and maximum load threshold respectively. In addition to the minimum and maximum threshold setting load, also need to set the largest threshold for four single load parameters. When a single load value of the nodes exceeds the maximum threshold, determine the node into the overload condition.

3.3 Select the target server based on the weighted probability

Set the current node value of real-time load percentage is ω_i, and P_i is probability value for the current geographic information service type request forwarded to nodes, the calculation formula for P_i as follows:

$$P_i = \frac{1-\omega_i}{\sum_{i=0}^{n-1}(1-\omega_i)}$$

Among them, n is the total number of servers in server cluster. The target server selection method based on the weighted probability can achieve better load balancing degree.

4 Experiment and analysis

4.1 The overall architecture of experimental system

In this paper, the designed load balancing algorithm was applied to a large geographic information service system of test environment. The experimental system is the service system of small building in the local area network (network bandwidth of 1000 MB/s) based on the test environment, which has two sets of data storage server and one set of metadata server data storage server cluster, mainly used for storage experiment data; Database server cluster is consists of two database server, mainly used for management the experimental data; Application server cluster is consists of eight sets of geographic information application server (late test process dynamically extension), four kinds of geographic information service mentioned above (map browsing, query analysis, data subscription and download, directories, and metadata service) are distributed for a server group of each category two servers, a total of four server group; One load balancer and two sets of Web server compose Web server cluster. Four PCS are used to simulate mass user concurrent access, specific structure as shown in Fig 3.

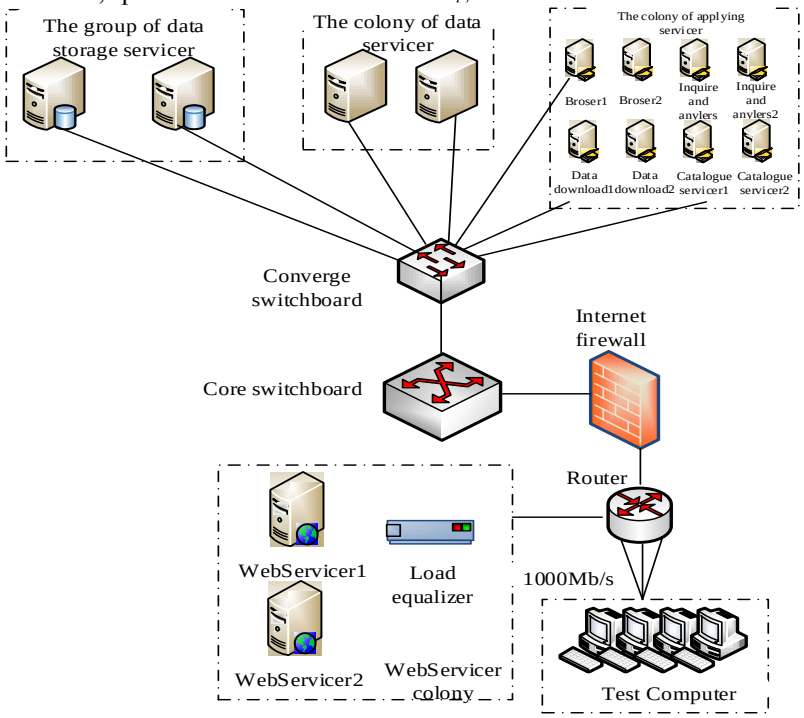

Fig. 3. General Architecture of Experimental System

4.2 The function and process of load balancing system

Load balancing system has realized load balancing algorithm proposed and intelligent scheduling of geographic information service, its function modules and process is shown in Fig 4, the load balancing system includes four modules: service receiving, dispatching decision module, load monitoring and load.

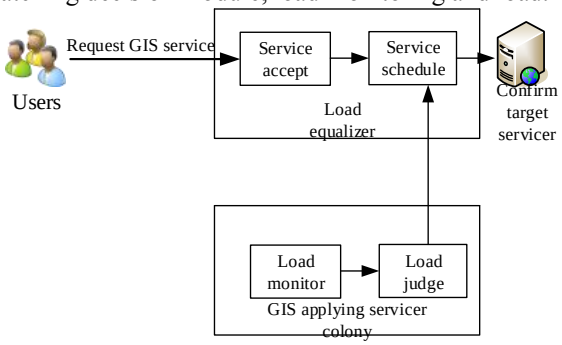

Fig. 4. Function and Process of Load Balancing System

Service receiving module is deployed on the load balancer, to complete geographic information services received, all user requests geographic information services are added to the task queue; Service scheduling module is deployed on the master-slave load balancer, mainly according to the above mentioned algorithm to determine the target server, to complete the scheduling of service; Load monitoring module is deployed in all geographic information application server cluster server, to complete the server load information collection; Load determination module is deployed in all geographic information application server cluster server, based on load information collected by the load query module, to finish load the determination of calculation and the load status of weighted value.

4.3 The experimental data and parameters

To choose a region within the scope of a 1:10 00000 all series scale vector data and 25 meters network of digital elevation model, used for query analysis and data download service, the area within the scope of a 1:10 00000 grade 1-16 map tiles data (generated by the area vector data), image tiles data and terrain tiles, used for map browsing service, as Experimental data, amount of data about a total of 10T.

According to the mentioned above, this designed load balancing algorithm needs to determine many parameters and thresholds, including geographical information service type the corresponding weight vector a_j, dynamic feedback cycle T, benchmark server processing power C_m, threshold L_{low} and L_{high}, four individual parameters corresponding to the maximum threshold. Through the extensive pressure experiment, determine the weights of components a_j as shown in Table 1:

Table 1. Weight Component of a_j

	cpu	*I / O*	*mem*	*net*
a_1	0.1	0.3	0.2	0.4
a_2	0.4	0.2	0.3	0.1
a_3	0.1	0.3	0.2	0.4
a_4	0.4	0.2	0.1	0.3

In this experiment, the configuration in the geographic information application server cluster is consistent, are all the two main frequency 1.9 GHz 6 core CPU, memory is 64G, therefore benchmark server processing ability C_m and the inherent processing capacity C_i are the same, then type (1) can be represented as follows:

$$\omega_i = a_j^{cpu} \times L_i^{cpu} + a_j^{I/O} \times L_i^{I/O} + a_j^{mem} \times L_i^{mem} + a_j^{net} \times L_i^{net}$$

Threshold value L_{low} and the value L_{high} are 0.5 and 0.9 respectively, four individual parameters corresponding to the maximum threshold is 0.95, T is 10 s, a value of k is 1.

4.4 The experimental process and results analysis

Large-scale concurrent access of users to simulate the client is installed on each test machine, and then start all testing machine, for roaming through operation at the same time for a particular region level 15 maps tiles, using the proposed algorithm, the RR algorithm and LARD algorithm respectively. Set the initial number of concurrent access is 20, every 10 seconds to repeat a concurrent access, a total of 5 minutes, calculate the average system response time and system throughput; On the basis of the above steps, increase the number of concurrent access to 40, 60, 80, 100, 120, 140, 160, 180, 200, calculate the average response time of the computing system;

Similar to map browsing service (static), the average response time on the basis of the above steps are tested under different number of concurrent access query analysis (in the case of slope analysis, static), data subscription and download (download the amount of data about 100 minutes), catalog and metadata service. Fig 5 (a), (b), (c) and (d) respectively are the system average response time for four kinds of geographic information services using three different algorithms under different number of concurrent access.

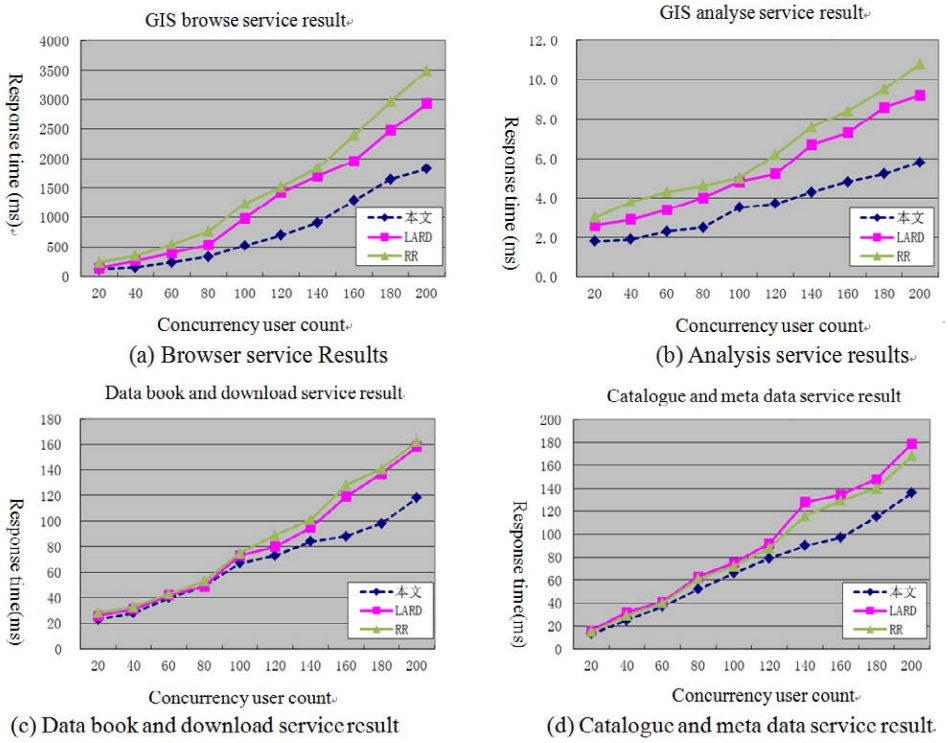

Fig. 5. Result of Experiment

We can draw the following conclusion by the Analysis diagram 5:

(1) With the increase of the number of concurrent users, the use of three kinds of load balancing algorithm of four kinds of geographic information service system response time are increased, that declare with the increase of concurrency server load pressure gradually increases, the system throughput increase gradually;

(2) For geographic information browse and query analysis services, under the same number of concurrent, this load balancing algorithm corresponding to the system response time is the shortest, and with the increase of concurrency this advantage is more obvious, then LARD algorithm, worst RR algorithm. Main reason is: geographic information browsing service in the form of access map tiles, and tiles and map is a map of a static cache. Because this design algorithm is one of the features of the geographic information service type to match up with the corresponding server group, which has improved the cache hit ratio, at the same time in the service dispatch considering the server's ability to work. Based on the RED algorithm can more accurately determine the server's load state, and in a certain period to dispatch service based on the weighted probability. Taken together, this algorithm can more accurately allocate the service request to the lighter load nodes, reduce the response time of the system.

(3) For download the data and metadata services, when concurrency is under 100, this algorithm’ s advantage is not obvious compared with other two algorithms, LARD and RR algorithm basic quite, the main reason is that the two types of service response time in addition to the related with the load balancing algorithm and the application server, to a large extent is related to the database server, when the concurrent user data has increased dramatically, load balance of the system response time impact gradually increased, this algorithm’ s advantage reflect gradually compared with the other two algorithms.

5 Conclusion

Load balancing algorithm is the core and key of intelligent dispatching for geographic information service. In order to develop with independent control, high reliability, high availability and extended geographic information service system, this paper studies a kind of suitable for network load balancing algorithm in geographic information service system, small geographic information service system is designed and the experiment results show that based on this algorithm can effectively realize the geographic information service resources reasonable scheduling, reduce system response time, can support multi-user high concurrent access to online geographic information service under the condition of operation.

References

National geographic information public service platform design guidelines[G] , Peking: SBSM, 2009.

Chen J, Jiang J, Zhou X, et al. Geographic information public service platform of technology on the overall design[J] , Geographic information world, 2009, No3: 7-11.

Chen J, Longgang Xiang, Jianya Gong. Network geographic information integration based on virtual earth shared service [J]， China science: Earth science, 2013, 43: 1770–1780.

Ziyou Wang, Minghui Zhou, Hong Mei. A dynamic client side load balancing mechanism[J]. China science: Earth science, 2013, 43: 60–72.

Genova Z, Christensen K J. Challenges in URL Switching for Implementing Globally Distributed Websites. Proceedings of The Workshops on Parallel Processing, 2000, 20(8): 89-94.

Guoqing Ren, Jinmin Yang, Dafang Zhang. Dynamic load balancing algorithm based on the content of the Web server[J]. Computer engineering, 2010, 36(13): 82-84.